

Basic Level Machine Learning Interview Questions

1. What is Machine Learning?

Machine Learning (ML) is a subset of Artificial Intelligence (AI) in which the algorithms are created, so that computers can learn, and make decisions without being explicitly programmed. It uses data to identify patterns and make predictions. For example, an ML algorithm could predict customer behaviour based on past data without being specifically programmed to do so.

2. What are the different types of Machine Learning?

Machine learning can be categorized into three main types based on how the model learns from data:

Supervised Learning: Involves training a model using labelled data, where the output is known. The model learns from the input-output pairs and makes prediction for unseen data.

Unsupervised Learning: Involves training a model using unlabeled data, where the system tries to find hidden patterns or groupings in the data.

Reinforcement Learning: Involves training an agent to make sequences of decisions by interacting with an environment, receiving feedback in the form of rewards or penalties, and learning to maximize cumulative rewards over time.

3. What is the difference between Supervised and Unsupervised Learning?

Supervised Learning: The model is trained on labelled data. Each training example includes an input and its corresponding correct output. The model's task is to learn the mapping between the input and output.

Example: Classifying the emails as spam or not spam.

Unsupervised Learning: The model is given unlabeled data and must find hidden structures or patterns in the data. No explicit output is provided.

Example: Clustering customers into different segments based on purchasing behaviour.

4. What is overfitting in Machine Learning?

Overfitting happens when a model learns both the actual patterns and the random noise in the training data. This makes it perform well on the training data but poorly on new, unseen data. Methods like L1/L2 regularization and cross-validation are commonly used to avoid overfitting.

5. What is underfitting in Machine Learning?

If a model is too simple to understand the patterns in the data, it's underfitting. This usually occurs if the model has too few features or is not complex enough. The model's poor performance is a consequence of its poor performance on the training and test data.

6. What is Cross-Validation?

Cross-validation is a method to check how well a machine learning model works. The data is divided into smaller groups called "folds." The model is trained on some folds and tested on others, and this is repeated for each fold. The results from all the folds are averaged to give a more reliable measure of the model's performance.

7. Explain the difference between Classification and Regression.

Classification: In classification problems, the aim is to predict a discrete label or category. The output is categorical, and models are used to assign the input data to one of these categories.

Example: Predicting whether an email is spam or not.

Regression: In regression problems, the aim is to predict a continuous value. The output is a real number, and models are used to estimate this value.

Example: Predicting the price of a house based on its features like size and location.

8. What is a Confusion Matrix?

A confusion matrix is a table used to evaluate how good a classification model is. The number of true positives, false positives, true negatives and false negatives is shown, useful for calculating performance metrics such as accuracy, precision, recall, and F1-score.

True Positive (TP): The positive class is correctly predicted by the model.

False Positive (FP): The model fails to predict the positive class.

True Negative (TN): The model predicts the negative class correctly.

False Negative (FN): The model gives the wrong answer to a negative class.

Accuracy:

Definition: Accuracy measures the proportion of correctly classified instances among the total instances.

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Precision (Positive Predictive Value)

Definition: Precision measures how many of the predicted positive instances are actually positive.

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP}$$

When to Use Precision?

Use precision when false positives (FP) are costly (e.g., spam detection, medical diagnosis).

Recall (Sensitivity or True Positive Rate)

Definition: Recall measures how many actual positive instances were correctly predicted.

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}$$

When to Use Recall?

Use recall when false negatives (FN) are costly (e.g., cancer detection, fraud detection).

F1-score (Harmonic Mean of Precision & Recall)

Definition: F1-score balances precision and recall into a single metric.

$$\text{F1} = 2 \times \text{Precision} \times \text{Recall} / \text{Precision} + \text{Recall}$$

When to Use F1-score?

Use when both FP and FN are important, especially in imbalanced datasets.

Metric	Best Used When	Formula
Accuracy	Dataset is balanced	$\text{TP} + \text{TN} / \text{TP} + \text{TN} + \text{FP} + \text{FN}$
Precision	False positives are costly	$\text{TP} / \text{TP} + \text{FP}$
Recall	False negatives are costly	$\text{TP} / \text{TP} + \text{FN}$
F1-score	Balance is needed	$2 \times \text{Precision} \times \text{Recall} / \text{Precision} + \text{Recall}$

9. What is an Activation Function in Neural Networks?

An activation function is a mathematical function applied to the output of a neuron in a neural network. It determines whether a neuron should be activated (i.e., fired) based on the weighted sum of its inputs.

Common activation functions include:

Sigmoid: Maps input to a value between 0 and 1.

ReLU (Rectified Linear Unit): Outputs 0 for negative inputs and the input itself for positive inputs.

Tanh: Maps input to values between -1 and 1.

10. What is Regularization in Machine Learning?

Regularization helps prevent overfitting by penalizing the loss function. The penalty discourages the model from fitting too closely to the noise in the training data. Common types of regularization include:

L1 regularization (Lasso): Adds the absolute values of the weights as a penalty term.

L2 regularization (Ridge): Adds the squared values of the weights as a penalty term.

11. What is Feature Scaling?

Feature scaling refers to the process of normalizing or standardizing the range of features in a dataset. This is essential when using algorithms that are sensitive to the scale of the data (e.g., gradient descent-based algorithms).

Common methods include:

Normalization: Rescaling features to a range between 0 and 1.

Standardization: Rescaling features so they have a mean of 0 and a standard deviation of 1.

12. What is Gradient Descent?

Gradient Descent is an optimization technique used to minimize the loss function in machine learning models. The model's parameters are updated with the negative gradient of the loss function. This update uses the learning rate to control how big the steps are.

Variants include:

Batch Gradient Descent: Uses the entire dataset to compute the gradient.

Stochastic Gradient Descent (SGD): Uses one data point at a time to update the parameters.

Mini-Batch Gradient Descent: Uses a small subset of the data for each update.

3. What is a Hyperparameter?

A hyperparameter is a variable that is set before learning starts. Hyperparameters control the training process and the model's architecture, such as the learning rate, the number of layers in a neural network, or the number of trees in a Random Forest.

14. What is a Training Dataset?

A training dataset is the data set used to train a machine learning model. It contains both the input features and the corresponding labels (in supervised learning). The model learns from this data by adjusting its parameters to minimize the error between its predictions and the actual labels.

15. What is K-Nearest Neighbors (KNN)?

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm. In KNN, the class of a data point is determined by the majority class of its k nearest neighbours. The "distance" between points is typically measured using Euclidean distance. KNN is a non-parametric algorithm, meaning it doesn't assume any underlying distribution of the data.

Intermediate Level Machine Learning Interview Questions and Answers

1. What is Dimensionality Reduction?

Dimensionality Reduction is the way of reducing the number of features (dimensions) in a dataset while retaining as much information as possible. It simplifies data visualization, reduces computational cost, and mitigates the curse of dimensionality. Popular techniques include:

Principal Component Analysis (PCA): Transforms features into uncorrelated components ranked by explained variance.

t-SNE: A visualization technique to map high-dimensional data into two or three dimensions.

2. What is Principal Component Analysis (PCA)?

PCA is a technique used for Dimensionality Reduction. It works by:

1. Standardizing the dataset to have a mean of zero and unit variance.
2. Calculating the covariance matrix of the features.

3. Identifying principal components by deriving eigenvalues and eigenvectors of the covariance matrix.
4. Projecting data onto the top principal components to reduce dimensions while retaining maximum variance.

3. What is the Curse of Dimensionality?

The Curse of Dimensionality means that working with high-dimensional data is challenging. As dimensions increase:

Data becomes sparse, making clustering and classification difficult.

Distance metrics lose significance.

Computational complexity grows exponentially. Dimensionality Reduction helps mitigate these issues.

4. What is Cross-Validation, and why is it important?

Cross-validation is a technique to assess model performance by dividing data into training and validation sets. The most common method is k-fold cross-validation:

- The data is split into k subsets (folds).
- The model is sequentially trained on a k-1 fold and validated on one fold. This ensures the model generalizes well to unseen data and avoids overfitting or underfitting.

5. Explain Support Vector Machines (SVM).

Support Vector Machine (SVM) is a supervised learning algorithm that supports classification and regression.

It works by:

- Maximizing the margin between different classes by finding a hyperplane.
- Using kernel functions (e.g., linear, polynomial, RBF) to handle non-linear data. SVM is effective in high-dimensional spaces and is robust against overfitting, especially in smaller datasets.

6. What is the Difference Between Bagging and Boosting?

- **Bagging (Bootstrap Aggregating):** Reduces variance by training multiple models on different bootstrapped datasets and averaging their predictions. Example: Random Forest.
- Boosting reduces bias by sequentially training models, each focusing on correcting the errors of its predecessor. An example is Gradient-Boosting Machines.

7. What is ROC-AUC?

The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds. The Area Under the Curve (AUC) measures the model's ability to distinguish between classes. A model with an AUC of 1 is perfect, while 0.5 indicates random guessing.

8. What is Data Leakage?

Data Leakage occurs when information from the test set is used during training, leading to overly optimistic performance estimates. Common causes include:

- Including target information in predictors.
- Improper feature engineering based on the entire dataset. Prevent leakage by isolating test data and strictly separating data preprocessing pipelines.

9. What is Batch Normalization?

Batch Normalization is a technique to improve deep learning model training by normalizing the inputs of each layer:

- It standardizes activations to have zero mean and unit variance within each mini-batch.
- It reduces internal covariate shifts, stabilizes training, and allows higher learning rates.

10. What are Decision Trees, and How Do They Work?

Decision Trees are supervised learning algorithms used for classification and regression. They split data recursively based on feature thresholds to minimize impurity (e.g., Gini Index, Entropy). Pros:

Easy to interpret.

- Handles non-linear relationships. Cons:
- Prone to overfitting (addressed by pruning or using ensemble methods).

11. What is Clustering, and Name Some Techniques?

An unsupervised learning technique for grouping similar data points is called clustering. Popular methods include:

- K-Means Clustering: Assigns data points to k clusters based on proximity to centroids.
- Hierarchical Clustering: Builds a dendrogram to group data hierarchically.
- DBSCAN: Groups based on density, identifying clusters of varying shapes and noise.

12. What is the Purpose of Feature Selection?

Feature Selection identifies the most relevant predictors to:

- Improve model performance.
- Reduce overfitting.
- Lower computational cost. Techniques include:
 - Filter Methods: Correlation, Chi-Square.
 - Wrapper Methods: Recursive Feature Elimination (RFE).
 - Embedded Methods: Feature importance from models like Random Forest.

13. What is the Grid Search Method?

Grid Search is a hyperparameter tuning method. It tests all possible combinations of hyperparameters to find the optimal set for model performance. For example, in an SVM:

- Search over kernels: Linear, Polynomial, RBF.
- Search over C values: {0.1, 1, 10}. Though computationally expensive, it ensures systematic exploration of hyperparameters.

14. What are the different types of Learning/Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables.

A. Supervised learning: [Target is present]

The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data.

The target variables are continuous linear regression, polynomial regression, and quadratic regression.

The target variable is categorical: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Gradient Boosting, ADA boosting, Bagging, Random forest, etc.

B. Unsupervised learning: [Target is absent]

The machine is trained on unlabeled data without any proper guidance. It automatically infers patterns and relationships in the data by creating clusters. The model learns through observations and deduced structures in the data.

Principal component Analysis, Factor analysis, Singular Value Decomposition, etc.

C. Reinforcement Learning:

The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

15. What is the difference between deep learning and machine learning?

Machine Learning:

Machine learning refers to algorithms that learn patterns from data without human programming. It uses a variety of models like decision trees, support vector machines, and linear regression to make predictions. ML typically works with structured data and requires feature engineering, where a human expert selects the features that are important for training the model.

Deep Learning:

Deep learning is a specialized subset of machine learning that uses artificial neural networks with many layers (hence “deep”). It can automatically learn features from raw data (e.g., images or text) without the need for manual feature extraction. Deep learning models are more computationally intensive and require larger datasets but are capable of achieving remarkable performance in tasks like image recognition, speech-to-text, and natural language processing.

Key Difference:

Deep learning models often outperform traditional machine learning models for tasks involving unstructured data (like images, video, and audio) because they can automatically learn hierarchical features from the data. However, deep learning requires more data and computational resources

16. How are covariance and correlation different from one another?

Covariance:

Covariance measures the degree to which two variables change together. If both variables increase together, the covariance is positive; if one increases while the other decreases, the

covariance is negative. However, covariance does not have a normalized scale, so its value can be hard to interpret.

Correlation:

Correlation is a normalized version of covariance, which measures the strength and direction of the relationship between two variables. It ranges from -1 to 1. A correlation of 1 means a perfect positive relationship, -1 means a perfect negative relationship, and 0 means no linear relationship. Correlation standardizes the covariance to make the relationship easier to interpret.

17. State the differences between causality and correlation.

Causality:

Causality refers to a cause-and-effect relationship between two variables. If variable

A causes variable B, then changes in A directly lead to changes in B. Establishing causality often requires controlled experiments or deep domain knowledge and is more complex to prove.

Correlation:

Correlation refers to the statistical relationship between two variables, meaning they tend to vary together, but it does not imply one causes the other. For example, there might be a correlation between ice cream sales and drowning incidents, but it does not mean that ice cream consumption causes drownings. It could be due to a third factor, such as hot weather.

Key Difference:

Causality establishes a direct cause-and-effect relationship, while correlation only suggests that two variables move together without implying causality.

18. What is Bias, Variance, and what do you mean by Bias-Variance Tradeoff?

They are both Errors in the Machine Learning Algorithms. This was just to say that when the algorithm can't really afford to generalize the right observation from the data, bias occurs. Now variance happens when the model overfits to small changes.

When building a model, if one starts adding more features, it will increase the complexity and we will lose on the bias but we gain some variance. This is a trade-off between bias and variance, in order to find the "perfect amount of error".

Bias:

Approximating real world problem with a simple model induces error which we call the bias. A high bias model relies heavily on the assumptions about the data, thus underfitting the data.

Variance:

Variance refers to the model's sensitivity to small fluctuations in the training data. A high-variance model may overfit the data, capturing noise or outliers instead of general patterns, leading to poor performance on unseen data.

Bias-Variance Tradeoff:

The bias-variance tradeoff is the balance between bias and variance. A model with high bias tends to underfit, while a model with high variance tends to overfit. The goal is to find a model that minimizes both bias and variance, resulting in the best generalization to unseen data.

19. What is Time Series?

A Time Series is a sequence of data points indexed or ordered by time. Time series data is typically collected at consistent intervals (e.g., hourly, daily, monthly) and is used for forecasting or identifying patterns over time. Time series analysis involves understanding trends, seasonality, and cyclical behavior to predict future values.

Example: Stock market prices, weather forecasting, and website traffic.

20. Explain the differences between Random Forest and Gradient Boosting machines.

Random Forest:

Random forest is considered an ensemble learning method that utilizes multiple decision trees trained on random subsets of the data. It uses bagging (Bootstrap Aggregating) to reduce variance by averaging the predictions of many trees. It works well for both classification and regression tasks and is robust against overfitting due to its random sampling.

Gradient Boosting Machines (GBM):

An ensemble method along the lines of Gradient Boosting is one that takes weak learners (usually decision trees) and improves their performance iteratively by building them sequentially. The loss function is minimized for each new tree, with errors from the previous ones. It is more prone to overfitting, but can also achieve better accuracy when tuned optimally.

Key Differences:

Training Method: Random Forest builds trees independently, while Gradient Boosting builds trees sequentially.

Overfitting: GBM is more prone to overfitting, but Random Forest is less so.

Performance: GBM typically provides better accuracy, but Random Forest is faster to train and easier to tune.

